Measuring Demonstrated Potential Domain Knowledge with Knowledge Graphs

Jiyin He
CWI, Amsterdam
jhe@cwi.nl

Marc Bron
Yahoo!, London
mbron@yahoo-inc.com

Abstract

Current search and recommendation engines enable us to effectively retrieve a set of documents based on topical relevance. What is not taken into account is the knowledge a user may already have about a topic, e.g., whether information is redundant or whether he/she is able to understand the results. We propose a method to measure demonstrated potential domain knowledge (DPDK) as a proxy for knowledge and use this metric to analyse the query log of a user spanning over 10 years.

1 Introduction

The Web consists of billions of web pages and the information contained within has the potential to fastforward technological developments, enable life long learning, and to solve many of our day-to-day problems. During a single lifetime, however, we are able to process only a fraction of these pages and as our capacity to process data is limited, carefully selecting the data that we do process becomes important. Current search engines enable us to effectively select a set of documents presented as a ranked list. However, these documents are selected based on topical relevance, i.e., the extent to which the document is about the query, and some contextual information about the user, i.e., age, location, and search history. What is not taken into account is the knowledge a user may already have about a topic and whether he/she is able to understand the results presented.[1] These factors seem especially relevant in situations where an information need goes beyond simple fact lookup, and a user is trying to master a new skill or learn about new topics. What is needed is a way to present users with documents, not only based on topical relevance, but also based on cognitive relevance, i.e., the relation between the user's state of knowledge and cognitive information need [9].

The cognitive relevance of documents depends on the individual user's current knowledge of, and assumptions about, the world. For example, an individual who has not studied mathematics beyond elementary arithmetic would not be capable of, or have great difficulty with, processing documents dealing with concepts of integral calculus [6]. Contrary to topical relevance, the cognitive relevance of documents to a user's information need is adapted to the users context, i.e., the user's knowledge of the world at a particular point in time. A consequence is that it is constantly changing as by processing new information a user's knowledge about the world may have changed.

The question is then, how to determine a user's knowledge. We argue that any such effort requires at least two components: expressions of a user's knowledge and a knowledge base. From work on the analysis of long term query logs we know that based on users' queries, changes in interest can be detected. For example, the query "mortgage" was found to be correlated with "calculator" and "lender" within a 30 minute session. In a window between one and seven days, however, this changed to terms such as "realtor" and "property." From one week to a month users used terms such as "insurance," "notary" and "IRS." Finally, between one and three months queries related to furnishing become more prevalent, e.g., "Bed Bath and Beyond" and "Pottery Barn" [7]. So if a sufficiently comprehensive log of a user's queries can be found, then

Copyright © by the paper's authors. Copying permitted for private and academic purposes.

In: L. Dietz, C. Xiong, E. Meij (eds.): Proceedings of the First Workshop on Knowledge Graphs and Semantics for Text Retrieval and Analysis (KG4IR), Tokyo, Japan, 11-Aug-2017, published at http://ceur-ws.org

[1] Although search engines may compile user profiles based on historical data these are at best implicit represenations of knowledge.

this may allow us to gain insights in a user's knowledge. Search logs alone, however, are not enough. They may be able to provide an overview of what a user does know, but it does not provide any information about what is unknown to the user, i.e., what a user could potentially learn about a particular topic. Knowledge bases are (often) manually constructed repositories aimed at providing an overview of the concepts that exist in a particular domain. These repositories could be seen as a representation of the knowledge of an expert in the domain.

Given these two components we propose a method to measure demonstrated potential domain knowledge (DPDK) as a proxy for knowledge and use this metric to analyse the query log of one of the authors spanning over 10 years. The choice to use a single log in our analysis was made for several reasons. The first is practical, i.e., it is hard to obtain a large sample of long term (e.g., over a decade) user logs. Second, it is difficult for an external assessor to judge what motivated users to issue particular queries. The third and final one is ethical, in our qualitative analysis we aim to understand the actions and motivations of the user over an extended period of their life and we feel that such an analysis should require user consent. We note that this study is exploratory in nature and understand the lack of generalizability of our findings beyond the log used in the study. However, we believe that our findings will motivate further study into this topic.

2 Methodology

Defining Knowledge. The definition of knowledge has been the topic of debate among scholars and philosophers since at least the ancient Greek times and is generally known as a branch of philosophy called epistemology. The work of the Greek philosopher Plato gave rise to an early definition of knowledge. In his work the Meno he puts forward the following paradox [8]: *For anything, F, either one knows F or one does not know F. If one knows F, then one cannot inquire about F. If one does not know F, then one cannot inquire about F. Therefore, for all F, one cannot inquire about F.*

In response Plato concedes that one cannot inquire about something one does not know, however, having a belief about something is adequate to start inquiry. Hence Plato describes different steps on the way towards knowing something:

- Perception of a phenomenon that stimulates forming a belief;
- to hold a belief about something and develop explanations for the belief;
- to verify that the belief is true and to finally know something.

Although Plato remains vague on how one actually transitions between stages his description of how one acquires knowledge suggests a definition of knowledge as a *justifiable true belief* (JTB):

Belief : something that is known must have been encountered, whether it is through perception or derivation.

True : the proposition, fact, or object that is believed must be true in order to be known.

Justifiable : a reason, explanation, or account that explains why someone holds a belief and believes that it is true.

This notion of justifiable true belief has a central place in epistemology. [5] More recent critics have suggested cases where a definition of knowledge as a justifiable true belief fails. Often referred to as Gettier problems or generalizations thereof. [4]. For example, the "cow in the field" problem, where a farmer checking up on his cow confuses a piece of black and white canvas caught up in a distant bush for a cow. However, since the animal actually is in the field, but lying hidden in a ditch the farmer has a justified, true belief, which does not seem to qualify as "knowledge" [2]. Others have suggested modifications to the theory of justified true beliefs by adding additional constraints or modifying the definition of justification to render such examples false.

In the remainder of the paper we focus on two necessary conditions for justifiable true beliefs, i.e., truth and belief.

Measuring Knowledge. One component in our discussion of epistemology above is *truth*, i.e., the objects or facts that exist and can be known. This is actually the object of study of another branch of philosophy, i.e., metaphysics. In his work Aristotle describes one of the first ontologies that provides a high level categorization of *the things there are*.[3] With the rise of the semantic web and Linking Open Data Cloud an ever increasing ontology or knowledge base is available that provides a reference for the facts and objects that exist in the world. It provides an excellent representation of the things in the world that someone can have knowledge about.

The second component is *belief*, i.e., in order to know an object someone should have a belief about that object. However, it is difficult to observe one's beliefs directly. Instead we can observe beliefs as expressed through interactions with digital systems, where users type queries, click on articles, and write messages. These expressions may serve as observable expressions of beliefs.

The final step is then to take the intersection between the truth (things existing in the world) and the user's beliefs. Recent advances in query understanding and entity recognition have resulted in systems that are able to reliably link user expressions with objects in knowledge bases [1]. For example, a user issuing the query "Michael Schumacher" could link this expression to the entry of Michael Schumacher the race car driver and derive related information from the assertions in the KB. This expression suggests that this user potentially knows who Michael Schumacher is and attempts to connect his belief with the truth through querying.

Having set the scene with the above definitions, we now explore quantities and their associated assumptions that can be used as metrics to operationalize the measurement of knowledge.

- First of all if we accept that the collective justifiable true beliefs held by a person constitute his/her knowledge, then counting the number of JTBs is a metric of knowledge.
- However we already relaxed the requirement for justification and only focus on counting TBs. Under this condition, if we assume that all beliefs of a user are observed as expressions and that all these expression are perfectly connected to an exhaustive KB of things then we measure the potential set of knowledge (TBs). A user's actual knowledge will be the subset of these that are justifiable (JTBs).
- Since no KB is complete, but if we assume that all beliefs of a user are observed and those that appear in the KB are successfully connected as true beliefs then we measure the user's potential knowledge within a certain domain as specified by the KB.
- Even if we observe all a user's expressions in digital systems during his/her lifetime, not all that user's beliefs will be observed. If both the observed beliefs and the KB are incomplete then we measure the potential observable beliefs of a user within a particular domain: one might call this "demonstrated/observed potential domain knowledge" (DPDK).
- The process of linking between observed expressions of beliefs and KBs is also not perfect. However, we treat this as measurement error.
- A user can have beliefs about objects (entities) but also relations between objects or aspects of objects. In this preliminary work we focus only on beliefs held about objects. We leave detecting beliefs about relations and aspects as well as how to link these to knowledge bases as future work.

Finally a note on forgetting. In the above argument we assume that the TBs of a person always increase as he/she acquires new information. We therefore treat forgetting as no longer being able to provide a justification for a belief. This seems reasonable since having observed that the belief was once held makes it a potential JTB and thus potential knowledge. We leave the handling of forgetting for future work.

3 Experiments and Results

Experimental setup and Data description. As our reference knowledge base of things that exist in the world we use DBpedia. DBpedia is a knowledge base extracted from Wikipedia and contains encyclopedic knowledge in the order of 580 million facts (relations) about 4.5 million objects (entities). In order to obtain expressions of beliefs we obtained the Google search query log of one of the authors from `https://takeout.google.com/settings/takeout`. The log ranges from May 2006 to May 2017 and contains about 62K queries. To link expressions of belief, i.e., queries, to the knowledge base we use a state of the art open source entity linking tool released by Yahoo! [1] [2]. If a query is successfully linked to one or more entities we select the entity with the highest log-likelihood score with a minimal score threshold of -3.0. Using this procedure we link 39K queries to entities in the knowledge base.

Figure 1 shows the number of queries per day, week, and month in darkgray as well as the number of entities linked from the queries to the knowledge base in lightgray. We observe that the number of queries issued has high variance, whether per day, week or month. There are times when many queries are issued and times when fewer queries are issued. Further there are some spikes in the number of queries issued suggesting high search activity in a particular day, week, or month. In the monthly data we observe that there is an increasing trend in the number of queries up to 100 months (end 2014) after which it drops. This aligns with the start of the author's use of an additional search engine. The number of entities linked follows the same trends as the number of queries issued, albeit at about 2/3 the volume. On average there are 15.57 queries per day, 107.0 queries per week, and 468.3 queries issued per month. The number of entities linked are on average 9.884 per day, 67.93 per week, and 297.3 per month. In terms of time periods without queries we find that in

[2] `https://github.com/yahoo/FEL`

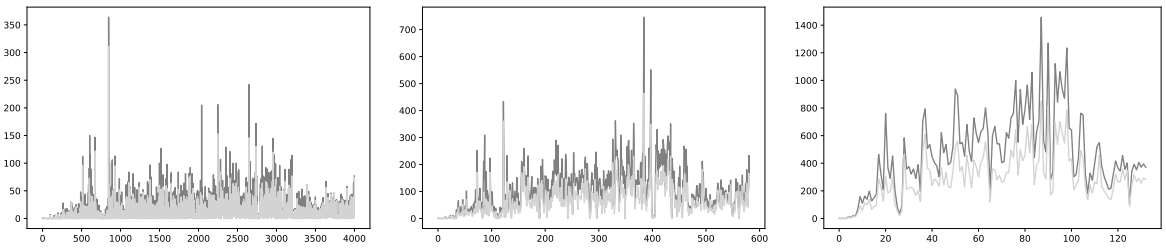


Figure 1: The number of queries per day, week, and month in darkgray as well as the number of entities linked from the queries to the knowledge base in lightgray. X-axis shows the number of days/weeks/months since May 2006.

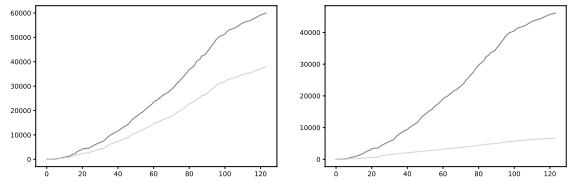


Figure 2: Left the cumulative number of queries (darkgray) and entities (lightgray) each month. Right the cumulative number of unique queries and entities, when only the first occurrence of a query (or entity) is taken.

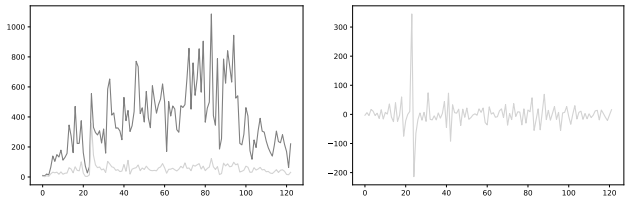


Figure 3: Left the derivative at each point of the cumulative unique queries (darkgray) and entities (lightgray), i.e., DPDK-velocity. Right the second derivative at each point of the cumulative unique entities, i.e., DPDK-acceleration.

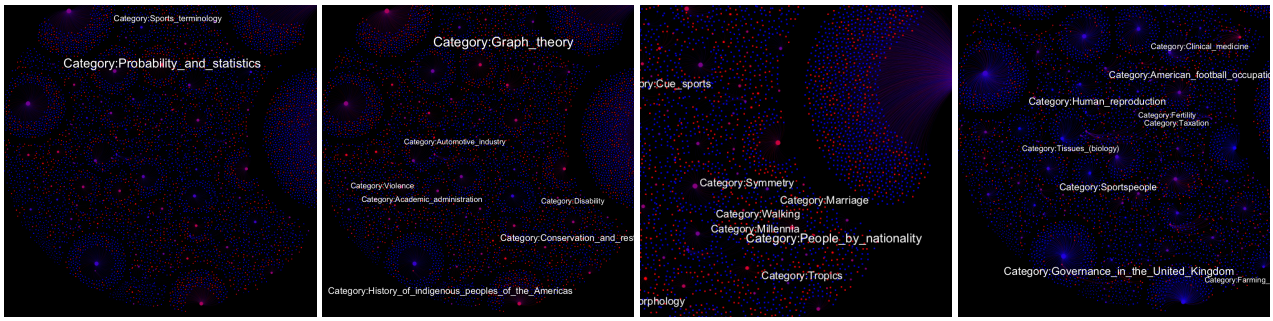
the period May 2006 to May 2017 there are 701 days without queries and 887 days on which no entities were linked. In terms of months we find that there are 3 months where no queries were issued, and 4 months in which no entities could be linked. These occurred all in 2006. Therefore, In the remainder of the paper, we will use the period Jan 2007 to Dec 2016 and analyze the data at a monthly granularity.

Quantitative Results. To measure demonstrated potential domain knowledge (DPDK) we compute the cumulative number of unique entities expressed per month. The entities are unique since only the first occurrence of an entity is considered. This reflects the intuition that multiple expressions of the same knowledge do not increase the overall DPDK. The left graph in Figure 2 shows the cumulative number of queries (darkgray) and entities (lightgray) each month. We observe that the cumulative number of queries and the cumulative number of entities follow a similar curve. This follows as well from Figure 1 where we observed that roughly 60 to 70 percent of the queries are consistently linked to an entity over time. The right graph in Figure 2 shows the cumulative number of unique queries and entities, when only the first occurrence of a query (or entity) is taken. Here we no longer observe similar trends between the unique cumulative queries and unique cumulative entities. The unique cumulative entities now exhibit a linear relationship over time. This is an interesting finding for two reasons. First, we observe that a metric based on queries and a metric based on entities measure different things. It suggests that the step of linking queries (or expressions of beliefs) to a knowledge base succeeds in differentiating between searching for information and expressing beliefs about facts in a domain. Second, it suggests that over 10 years time the author sought out new knowledge at roughly the same rate. It would be interesting to investigate whether this pattern generalizes to a wider sample and whether it correlates with curiosity or other personality traits.

Next we examine two derivative metrics from the DPDK, that is DPDK-velocity and DPDK-acceleration. The left graph in Figure 3 shows the derivative at each point of the cumulative unique queries (darkgray) and entities (lightgray). We observe that the number of unique queries per month has high variance similar as observed for the number of queries per month in Figure 1. In contrast the number of unique entities per month is more stable, with one exception around month 25. This is more clearly visible in the right graph that shows the second derivative at each point of the cumulative unique entities (for clarity the analogous data for queries is not shown). The author suspects this spike is due to finishing a MSc thesis followed by a long holiday.

This part shows that using unique entities as a basis for measuring beliefs is different from using queries. It suggests that query volume is not necessarily driven by quests for new knowledge but also by re-finding or perhaps knowledge outside of the domain of Wikipedia. We have not touched on what kind of potential knowledge was expressed by the author. We look into this next.

Qualitative Results. The above analysis treats all expressions that are linked to the knowledge base as observed expressions of beliefs about the domain of the knowledge base. It is not uncommon, however, for sub-domains to exist within a



(a) 2012-2013 (b) 2013-2014 (c) 2014-2015 (d) 2015-2016

Figure 4: The concepts linked from queries issued within a particular year as vertices in red and concepts linked from queries issued within other years as vertices in blue. Edges represent has_category relations and clusters are formed by concepts that are all of the same category. The central vertex in each cluster is the category, and the names of the top k categories are displayed.

knowledge base, especially one as broad as DBPedia. People, then, may express beliefs about a diverse set of domains during a particular time, or such expressions could be focussed on a particular domain. To analyze whether we observe such behavior in the author's log, we utilize the DBPedia categories, representing different domains, to cluster queries that are linked to entities within the same category in the knowledge base.

The Wikipedia category structure is not a strict hierarchy. To assign each entity to a single category at a particular level in the category structure we extract a hierarchy from the category structure. We use `Category:Main\_topic\_classifications` as the root node of the category hierarchy, eliminate cycles, and pick the shortest one to the root when multiple paths exist. Given this hierarchy, we find for each entity all its categories and the shortest path to the root. We can then slice the hierarchy at a particular level and find for each category all the entities that are associated with it. The amounts of identified entities covered under different categories can be seen as a distribution of a user's potential domain knowledge. One way to use this distribution, for instance, is to compare the knowledge distribution between different people, e.g. experts vs. novices. Since we lack the logs of multiple people, here we inspect changes in potential domain knowledge within the same log year over year.

Figure 4 shows the concepts linked from queries issued within a particular year as vertices in red and concepts linked from queries issued within other years as vertices in blue. Edges represent has_category relations and clusters are formed by concepts that are all of the same category. The central vertex in each cluster is the category name. The visualization was created using the open source tool Gephi[3] and enables exploration of how the number of concepts a user potentially knows increases over time. For clarity we only show the names of the categories with the highest acceleration of concepts linked from expressions during a year in a particular domain.

We observe that in 2012-2013 most expressions were linked to the category "Probability and Statistics". These queries were issued during the author's time as a PhD student in Information Retrieval. The author suspects that this may have prompted an increase in queries related to that category. In 2013-2014, after just finishing up as a PhD, the author was teaching a course on Social Network Analysis to undergraduate students and suspects that is the cause of an increase in the proportion of queries related to the category "Graph Theory". In 2014-2015 it is more difficult to single out a category in which new knowledge was demonstrated, but the authors' note that they were preparing their wedding in that year, explaining the rise in queries related to the category "Marriage". The visualization of 2015-2016 shows a number of interesting categories "Human Reproduction", "Fertility", and "Tissues_(Biology)" as well as "Governance_in_the_United_Kingdom and Taxation." These categories again align with some of the life events of the authors' during that year, i.e,. expecting their first child and moving to the United Kingdom.

The kind of qualitative analysis of a query log, like the one above, would be very difficult for anyone but the one who issued the queries. Even then it is hard to know what prompted an acceleration in queries related to concepts in a particular domain in hindsight. What we do observe is that some of the accelerations can be explained by the occurrence of certain life events. An interesting direction for future work would be to analyze how accelerations observed at different (lower) levels of the hierarchy, i.e., more specialized domains, relate to events in users' lives or more specific tasks.

[3] `https://gephi.org`

4 Discussion and Conclusions

In this paper we explored the potential of using search logs and knowledge bases to gain insight in a user's potential knowledge of certain domains. We explored the definition of a metric of knowledge based on the theory of justifiable true beliefs. We further operationalized the measure as demonstrated potential domain knowledge (DPDK), i.e., a person's observed potential beliefs within a particular domain, based on a series assumptions that relax the requirements for JTBs. We showed that measuring the number of expressions of beliefs (e.g., queries) that can be linked to a knowledge base is different from measuring the number of queries issued by a user. The former was observed to, surprisingly, increase linearly over time. Further we found that changes in DPDK within sub-domains of a knowledge base can be associated with certain life events of a user. Although anecdotal in nature these observation suggest that DPDK captures situations in which a user is learning about new things and increasing his/her knowledge.

Obviously this is the first step in creating a metric for demonstrated potential domain knowledge. We briefly highlight some promising directions for future work: (i) to validate the score with user experiments; (ii) to improve the accuracy of the metric by improving the linking of expressions to knowledge bases; (iii) linking expressions to relations as well as objects; (iv) using the score in applications such as ranking with cognitive relevance or improving targeting for advertisements and articles; and (v) extending the score to also incorporate justifications of user's true beliefs.

Acknowledgments

This research was supported by the Netherlands Organisation for Scientific Research (NWO) under project nr 13675.

References

[1] R. Blanco, G. Ottaviano, and E. Meij. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 179–188. ACM, 2015.

[2] M. Cohen. *101 philosophy problems*. Routledge, 2013.

[3] S. M. Cohen. Aristotle's metaphysics. 2000.

[4] E. L. Gettier. Is justified true belief knowledge? *analysis*, 23(6):121–123, 1963.

[5] A. I. Goldman. What is justified belief? In *Justification and knowledge*, pages 1–23. Springer, 1979.

[6] S. P. Harter. Psychological relevance and information science. *Journal of the American Society for information Science*, 43(9):602, 1992.

[7] M. Richardson. Learning about the world through long-term query logs. *ACM Transactions on the Web (TWEB)*, 2 (4):21, 2008.

[8] A. Silverman. Plato's middle period metaphysics and epistemology. 2003.

[9] A. Spink, H. Greisdorf, and J. Bateman. From highly relevant to not relevant: examining different regions of relevance. *Information Processing & Management*, 34(5):599–621, 1998.